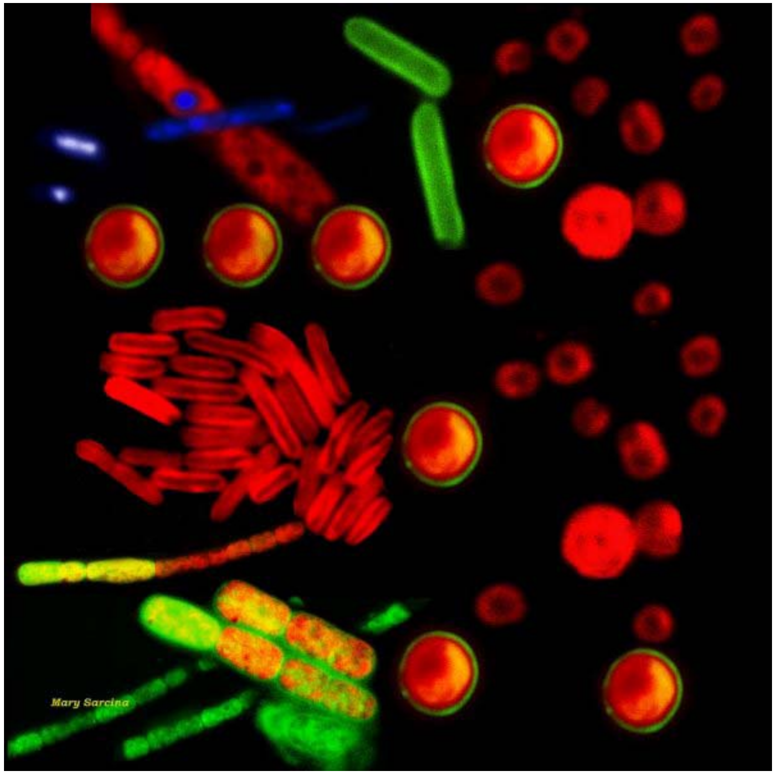
Biology and High-Performance Computing

Rick Stevens
Argonne National Laboratory
University of Chicago

Outline

- Trends impacting biology
- Survey of agency bio initiatives in the US
- The Scientific Opportunities
- Examples of Current Projects
- Future Vision
- Architecture Requirements
- Grids and Biology
- Conclusions

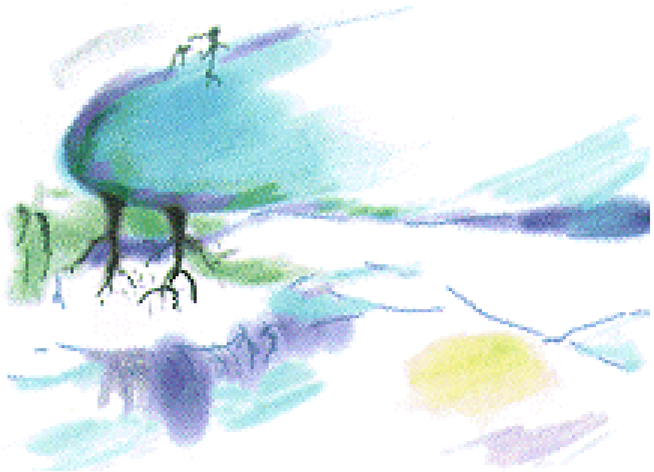


Trends Impacting Biology and HPC

- Increased availability of high-throughput technologies
  - Genomics, Proteomics, Imaging, etc.
- Rise of Bioinformatics as a discipline
  - Research programs and educational programs
  - Significant interest from Computer Scientists
- Emergence of Systems Biology
  - New foundation for Theoretical Biology
- Inexpensive and available computing resources
  - PC Clusters and Grids
- Increased awareness of the impact of HPC
  - 800 pound gorilla's are waking up (pharmas and NIH)

Survey of Recent Bio Initiatives in the US

- NSF
  - Biocomplexity and Tree of Life Initiatives
- NIH
  - Biological Science and Information Technology Initiative (BISTI)
  - Genomics and Structural Biology Initiatives
  - Biodefense Initiatives (RCE's etc.)
- DOE
  - Genomes to Life Initiative
  - Structural Biology Initiative
- DARPA
  - BioSpice Program
  - Biodefense programs
- NASA
  - Astrobiology Program



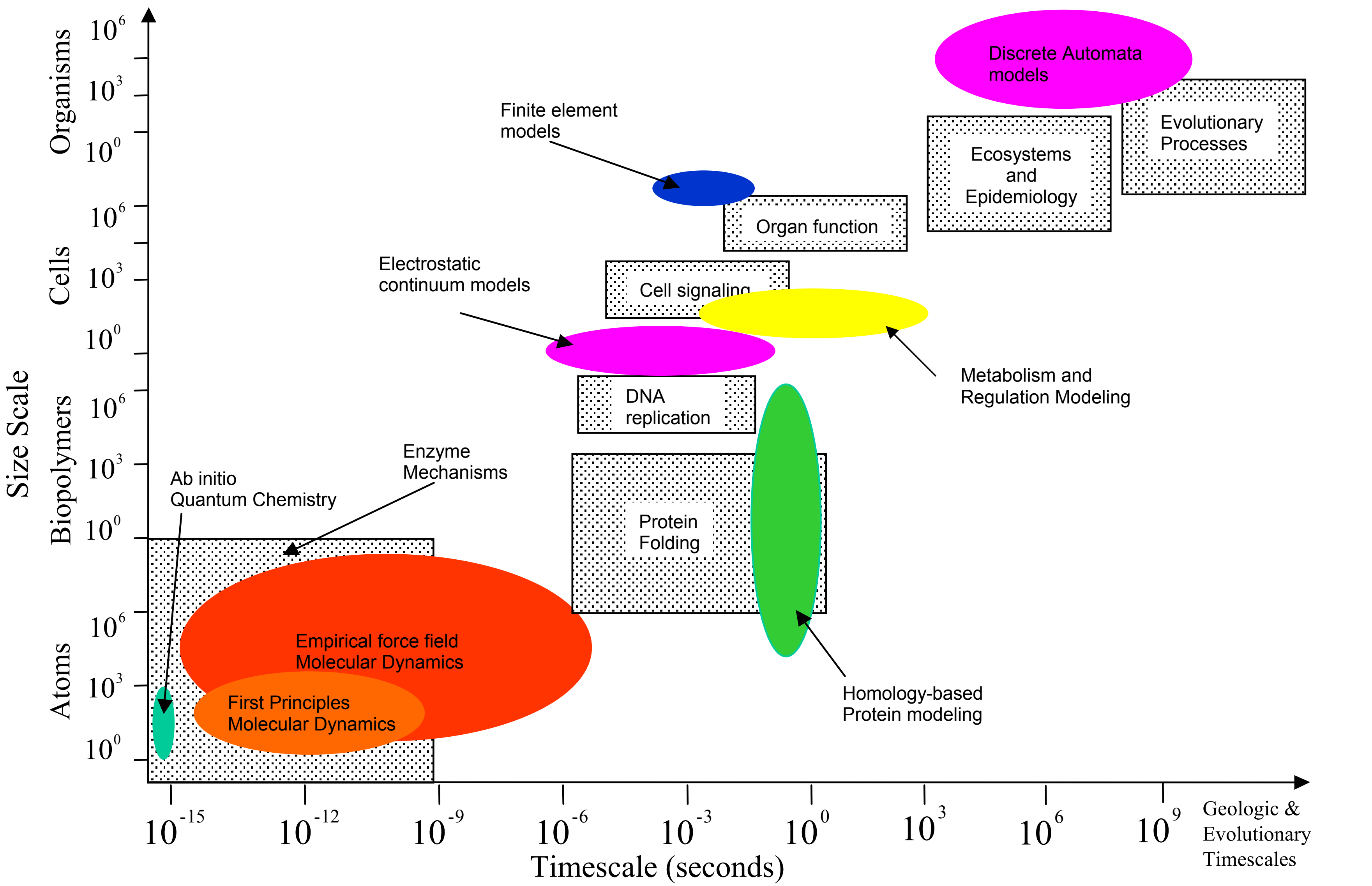
A Few Of The Scientific Opportunities

- First principles modeling of biomolecular systems
  - Protein folding, complexes, ion channels, reaction kinetics
  - Will continue to be a major driver for HPC
- From genome annotation to reverse engineering of biological systems
  - Metabolic network reconstruction
  - Regulatory network analysis
  - Populations and ecosystems
- From computational molecular biology to computational cell biology
  - Whole cell modeling (bacteria to human cells)
  - Tissue and organ modeling (hearts and brains)
  - Virtual organisms (microbes, worms, flies, plants, trees, mice and humans)
- Bioengineering
  - Structural, electrical and physiological models
  - Personalized response to therapy

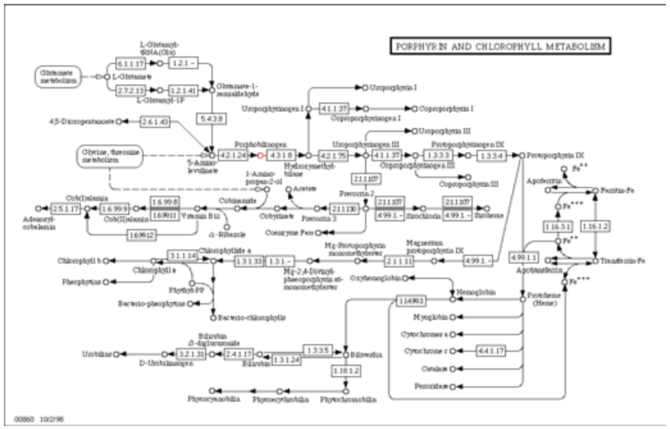
The Science for the 21st Century?

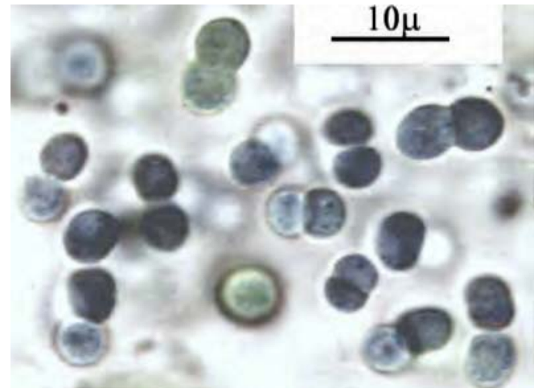
- The application of advanced biological thought and related technology could yield:
  - Safe and abundant food supplies
  - Sustainable and benign energy sources
  - Effective management of disease and aging
  - Novel materials and renewable industrial feedstocks
  - Advanced computational devices beyond Moore’s law
  - Wide variety of molecular scale machinery
  - Self-assembly and self-reproduction technologies

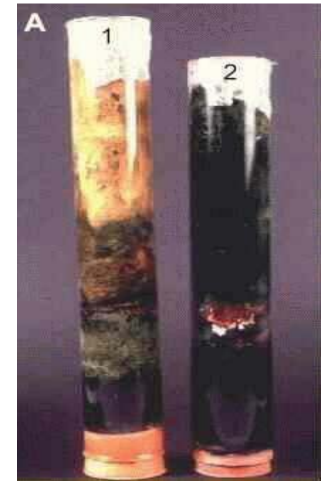
24 Orders Magnitude of Spatial and Temporal Range

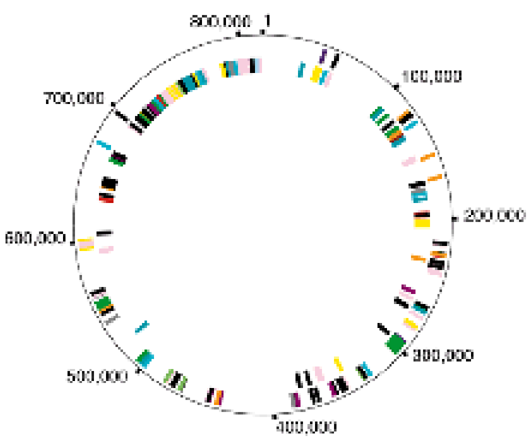


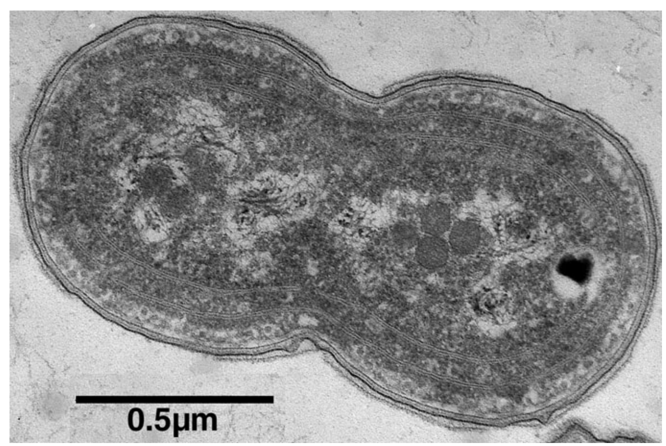
Genes → Cell Networks → Populations



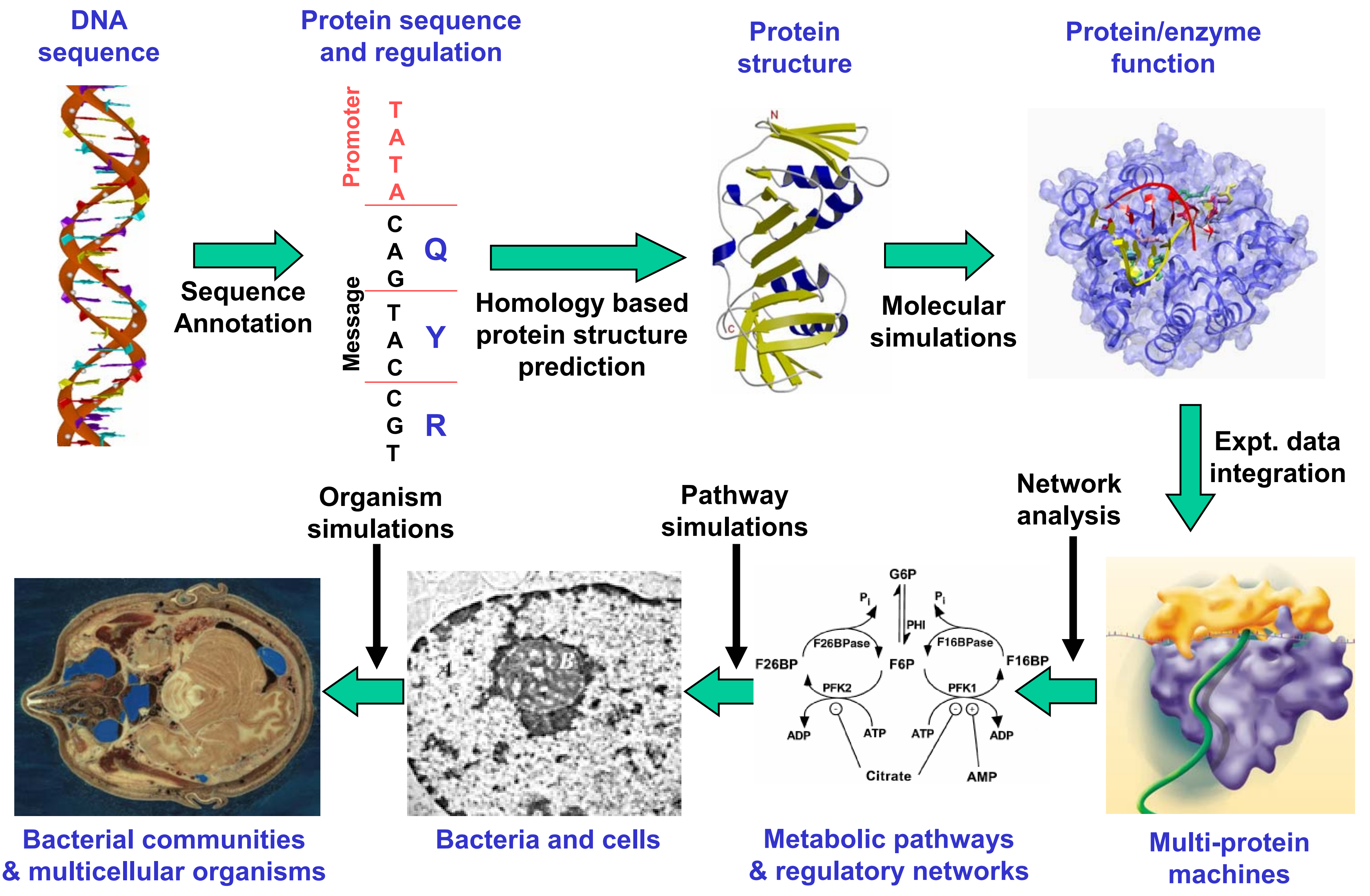








Computational analysis and simulation have important roles in the study of each step in the hierarchy of biological function

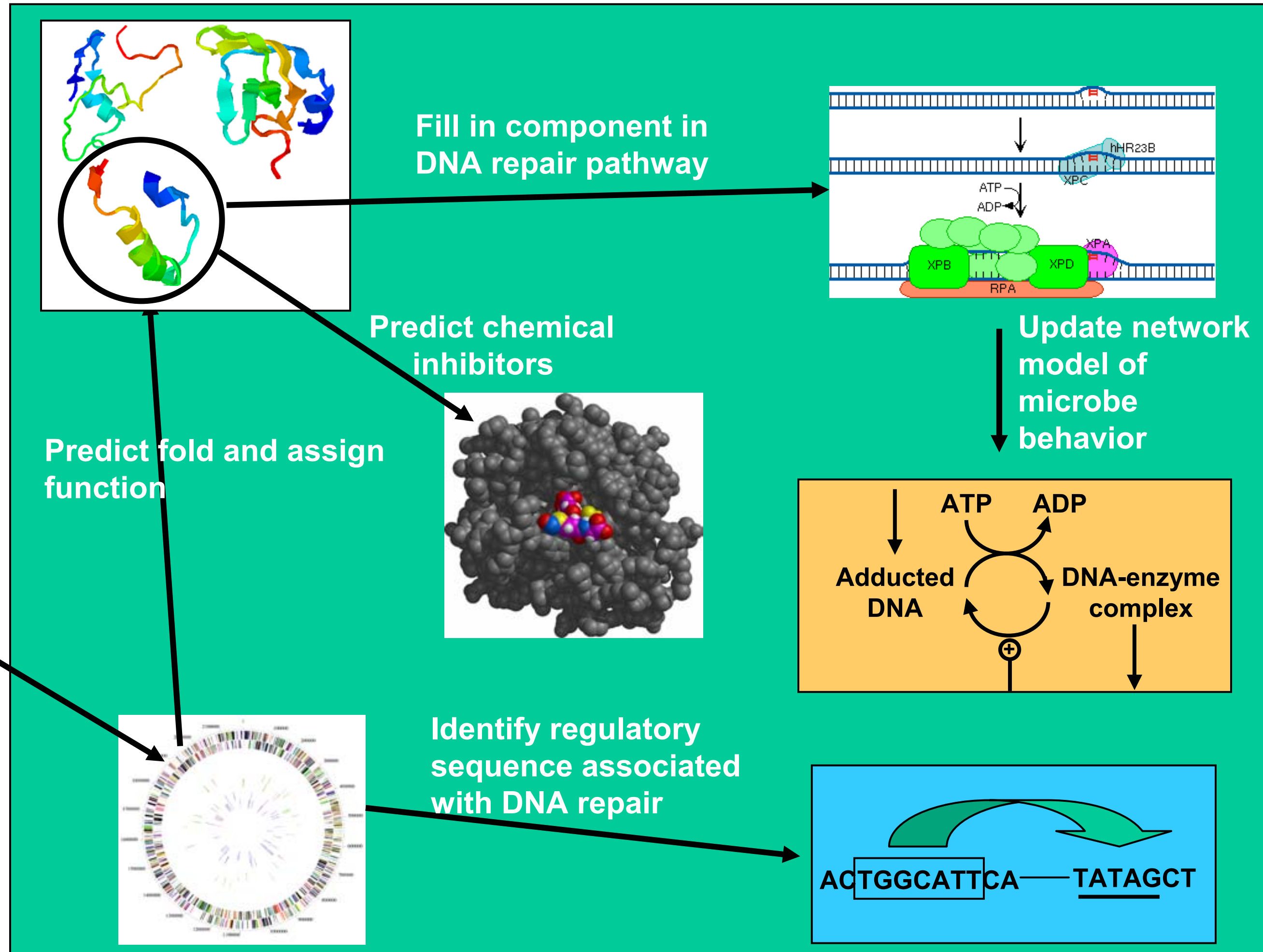


The New Biology in Action

Hypothetical example:

New data showing that a gene forms complex with DNA repair enzyme

Add function to genome database



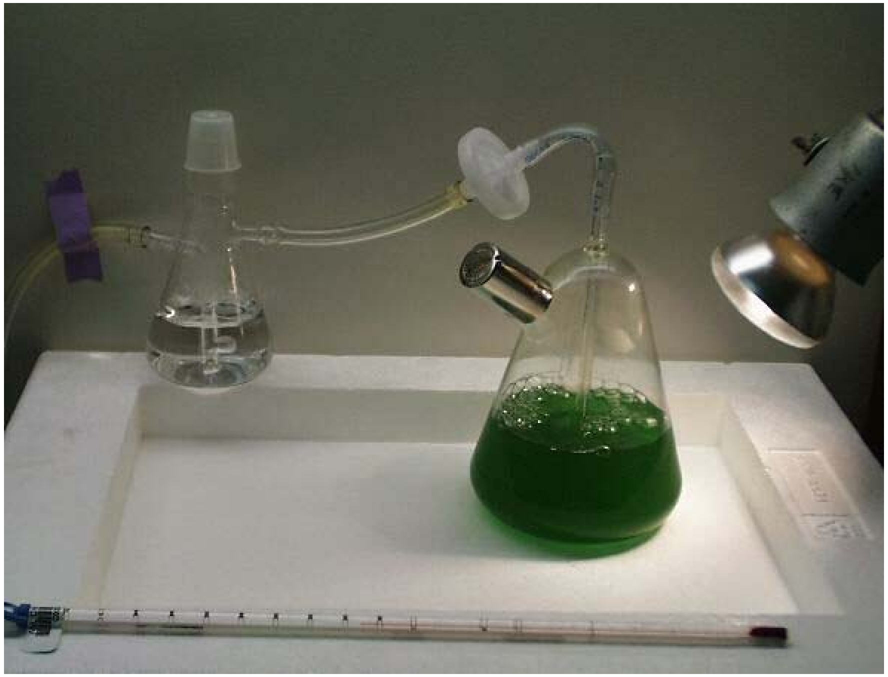
From Mike Colvin, LLNL

Towards a Systems Biology

- Integrative understanding of a biological system
  - Cell, organism, community and ecosystem
- Counterpoint to reductionism
  - Requires synthesizing knowledge from multiple levels of the system
- Discovery oriented not necessarily hypothesis driven
  - Data mining vs theorem proving

Example Problem: Circadian Clock

- Understanding the circadian clock in cyanobacteria
  - Ubiquitous problem in biology
- The proposed theoretical mechanism involves feedback and dynamics of core metabolism rather than oscillations in a genetic regulatory circuit
  - Community is split on this issue
  - Can’t be answered with just experimental data
    - Need modeling and simulation
    - Need supporting experiments
- Complementary wet lab work is relatively accessible



Circadian Clocks

- Must be stable over temperature ranges
- Must be stable through multiple generations (progeny are created with clocks in phase)
- Must be stable in response to noise
- Used by the organism to coordinate processes
  - Separation of incompatible biological processes
  - Cell cycle timing
  - Control variations of physiological state (e.g. buoyancy in pelagic species)
- Higher-organisms have centralized clocks (e.g. in one extremely limited region of the mammalian brain (the suprachiasmatic nucleus or SCN), there are groups of 16-thousand cells that act as master clocks) and secondary clocks distributed throughout

Temporal Separation Of Two Incompatible Cellular Processes

Left: During the light phase glycogen (●) goes up, oxygen evolution (not shown) is high, and N_2 fixation (○) is shut down. _Right:_ The same experiment under continuous dark. The is a clear proof that the external light modulation is not the primary cause of the temporal organization. (From Schneegurt M.A., et al. _J.Physiol._, 33, 639-642)

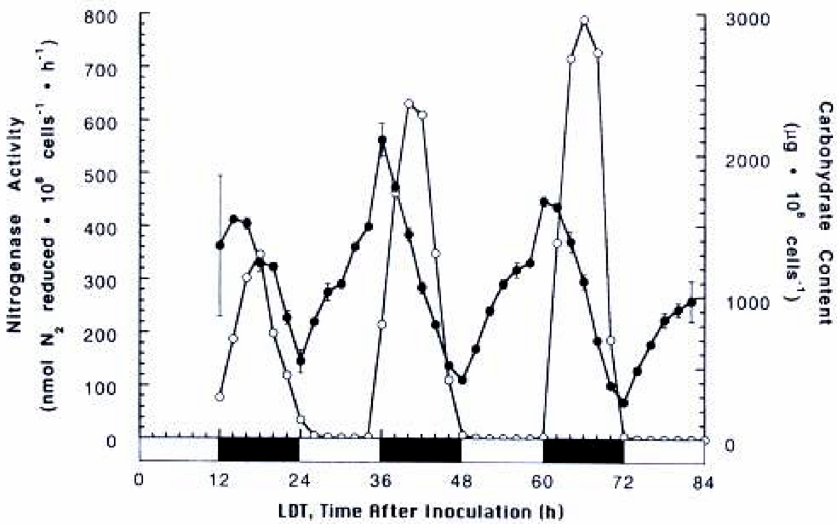


FIG. 5. Oscillation of N_2 fixation and carbohydrate content in *Cyanothece* sp. strain MGD grown mixotrophically with 50 mM glycerol under 12-h light/12-h dark (LD) conditions. Nitrogenase activity (○) was measured as acetylene reduction and the averages of duplicate assays are presented. Carbohydrate content (●) was assayed using anthrone reagent and the averages of three assays ± SD are presented.

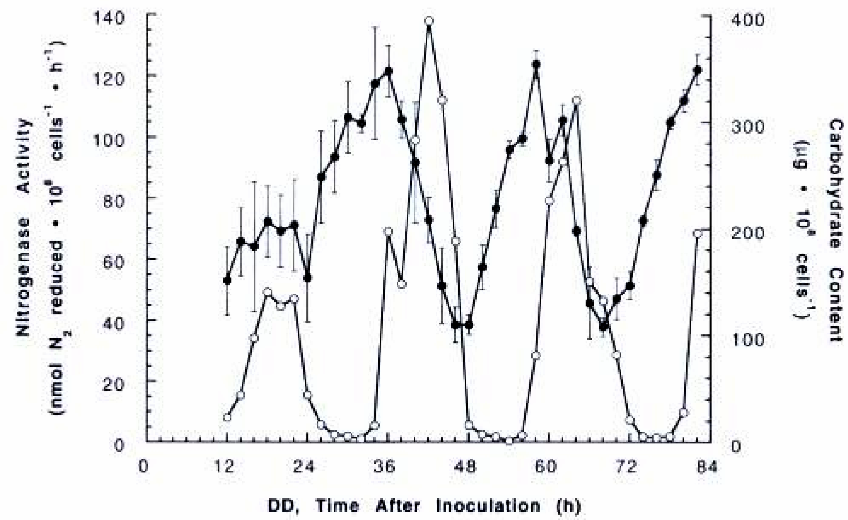


FIG. 6. Oscillation of N_2 fixation and carbohydrate content in *Cyanothece* sp. strain CGD grown chemoheterotrophically with 50 mM glycerol under continuously dark (DD) conditions. Nitrogenase activity (○) was measured as acetylene reduction and the averages of duplicate assays are presented. Carbohydrate content (●) was assayed using anthrone reagent and the averages of three assays ± SD are presented.

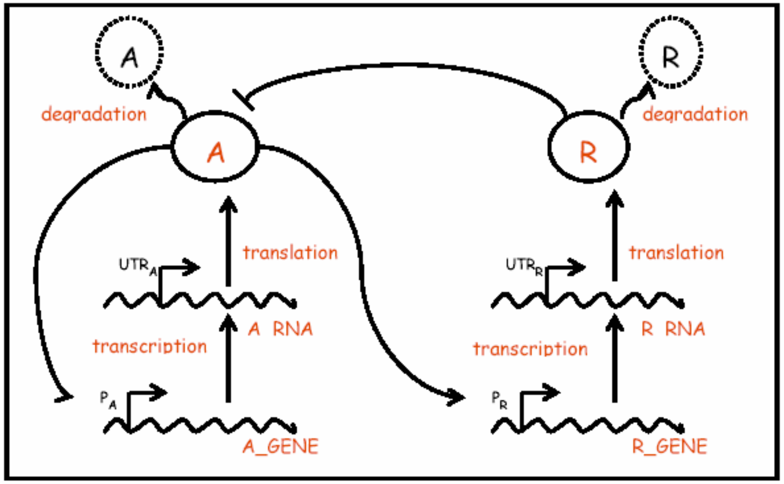
In Search of Fundamental Clock Mechanism

- Mechanisms not yet well understood
- Transcriptional regulation vs. post-transcription regulation vs. metabolic dynamics
  - Which process drives the other?
  - How to attack the problem?

"The evidence indicates that the clock network is based predominantly on transcriptional regulation…

The ability to maintain constant circadian periodicity despite global changes in the state of the cell is probably necessary for the circadian clock to be successfully embedded within the cell…

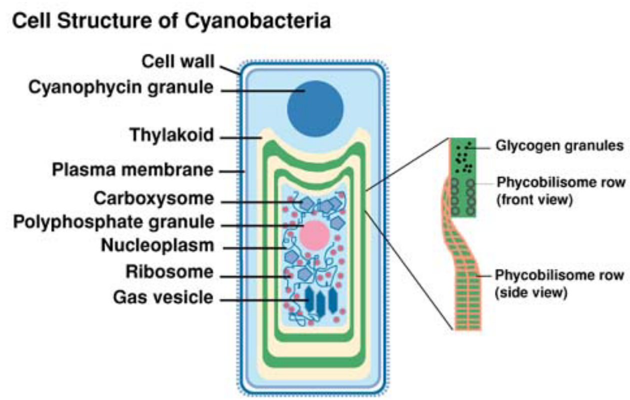
It is not clear whether this hysteresis-based network is the mechanism underlying circadian oscillations…"

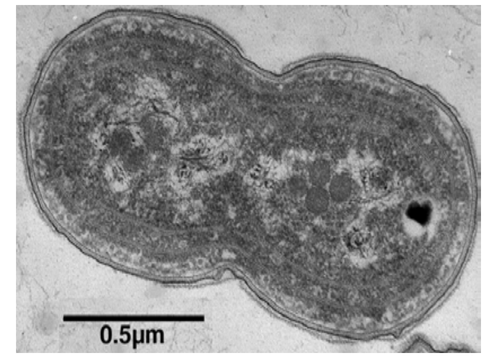


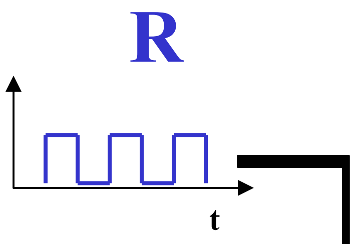
[Barkai N. and Leibler S. (2000) Biological rhythms: Circadian clocks limited by noise, Nature 403, 267 - 268.]

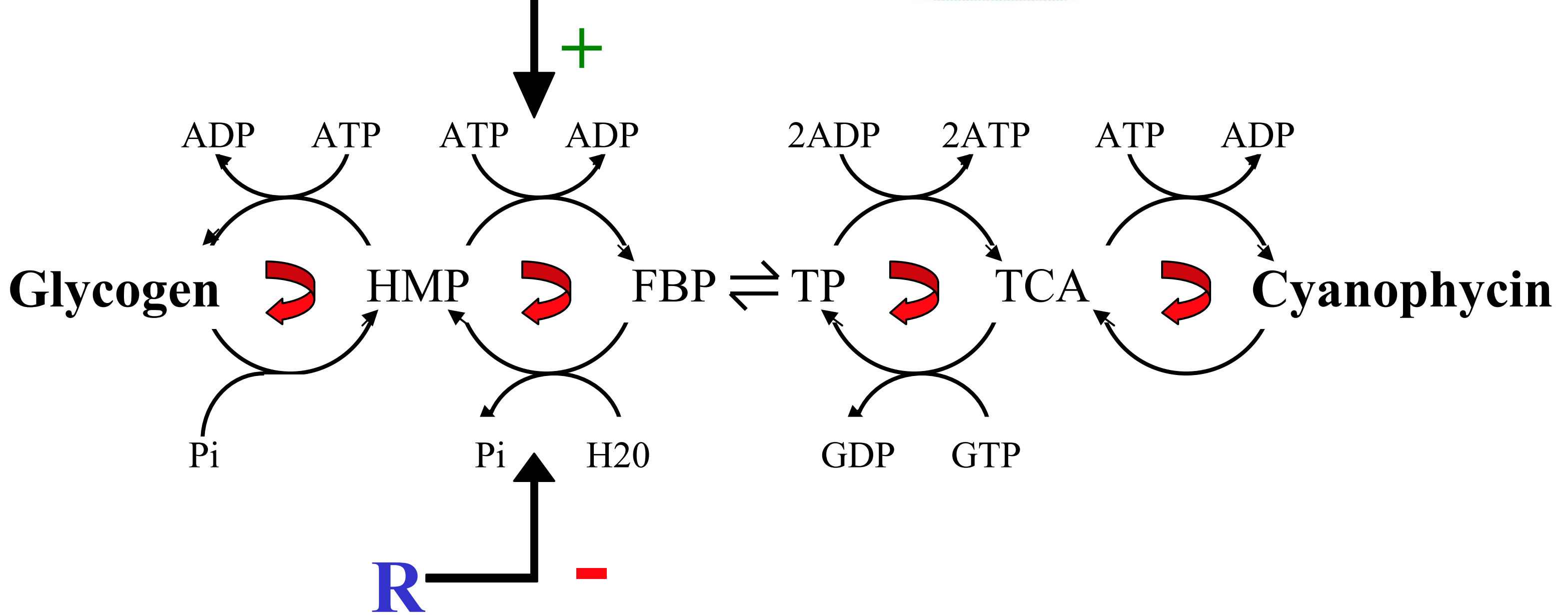
Substrate Cycles Controlled In Time

- Regulation results in anti-phase oscillations in glycogen and cyanophycin content
- Possible basis for circadian clock





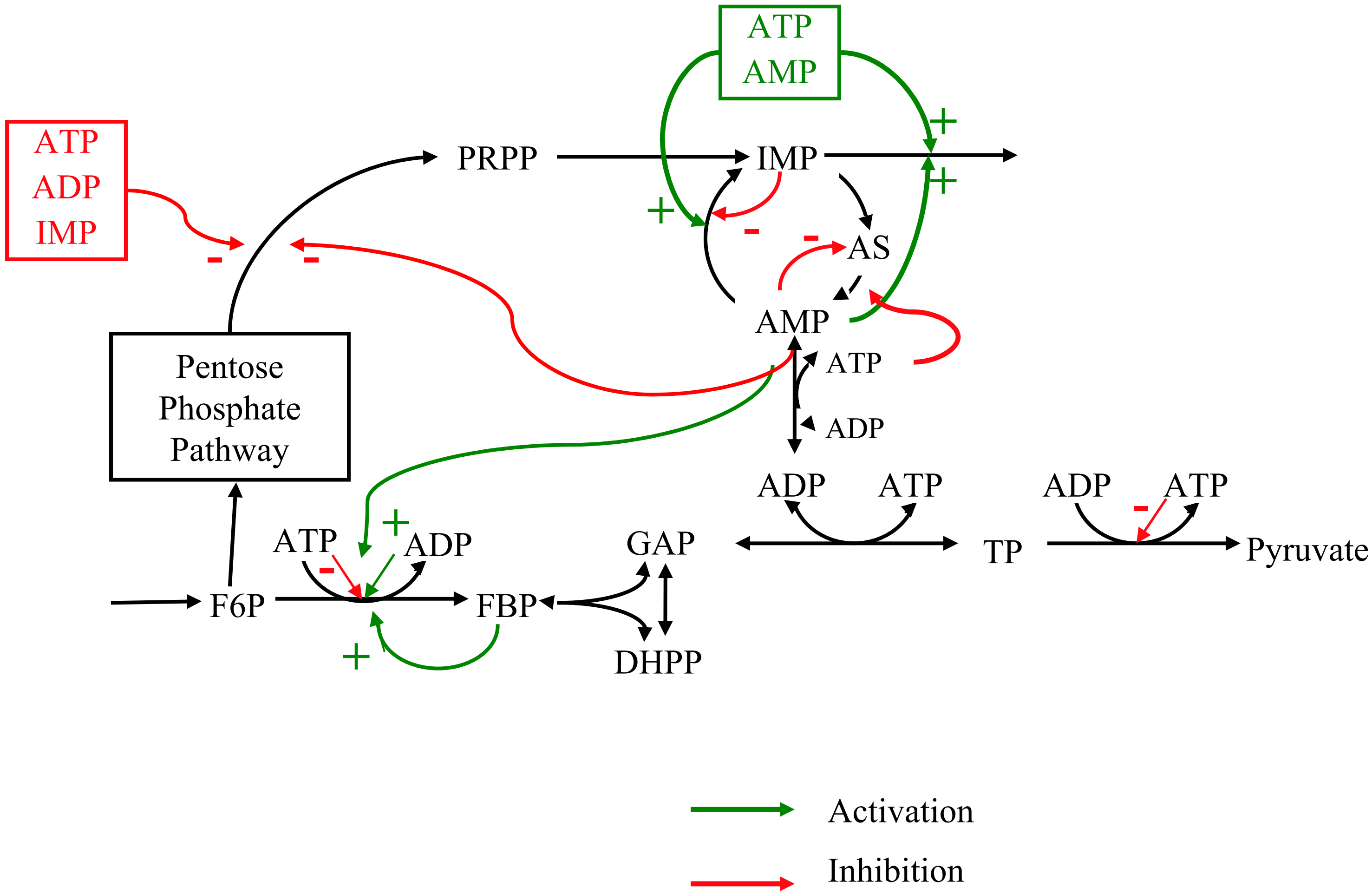




Futile substrate cycling

Beyond Sequence Analysis

Theoretically predicted allosteric regulatory mechanisms via optimization of a mathematical model derived from metabolic reconstruction

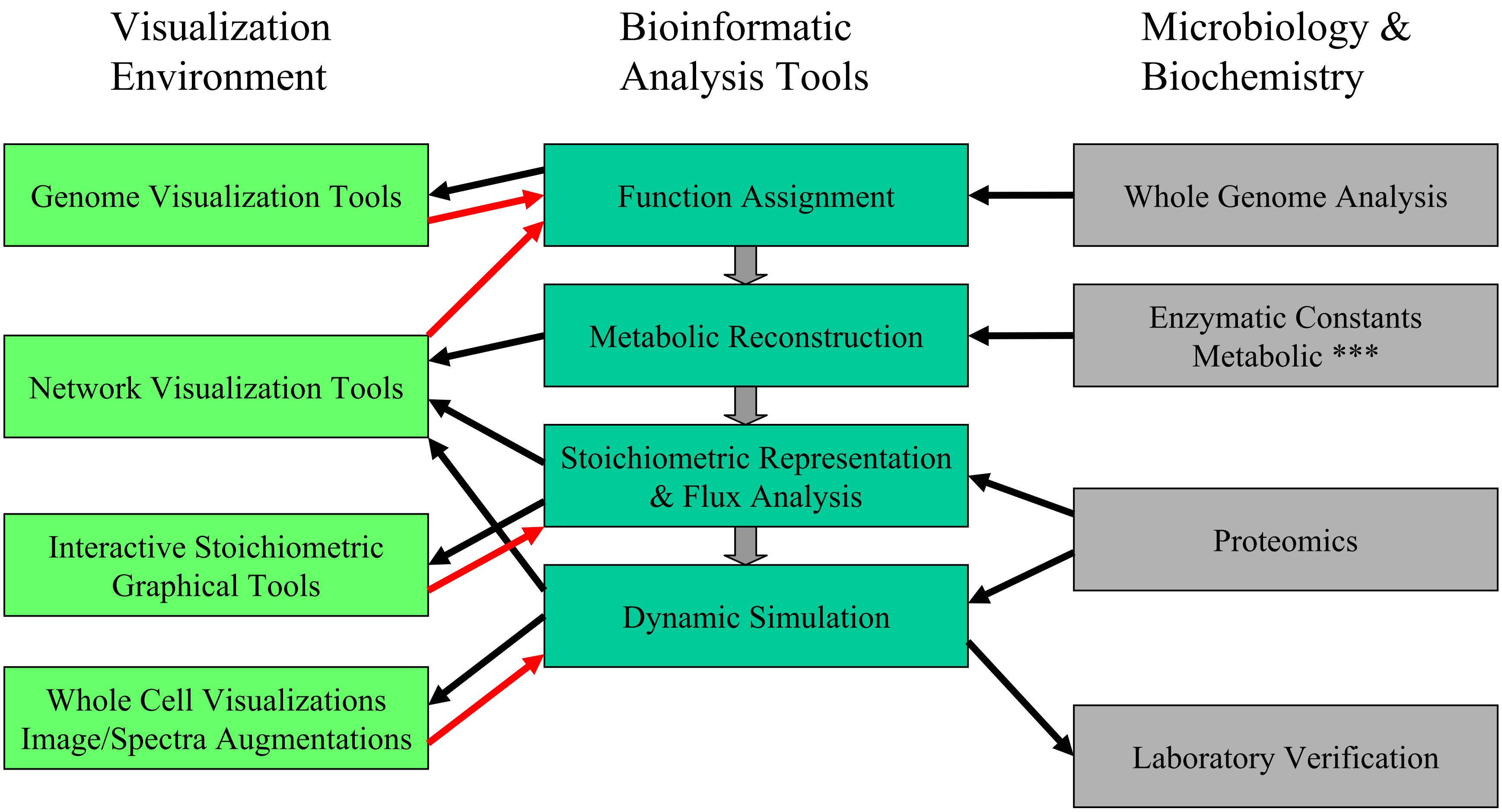


Systems Biology Model Development

Systems	Director	Institution	Features
ERATO/SBW,j	John Doyle	Caltech	planned workbench
Gepasi,w	Pedro Mendes	Santa Fe	MCA, systems kinetics
JarnacScamp,wx	Herbert Sauro	Caltech	MCA, Stochastic
StochSim,w+	Dennis Bray	Cambridge	Stochastic
BioSpice,u	Adam Arkin	LBL	Stochastic
DBSolve,w	Igor Goryanin	Glaxo	enzyme/receptor-ligand
E-Cell,u+	Masaru Tomita	Keio	metabolism. Net ODE
Vcell,j	Jim Schaff	U.CT	geometry
Xsim,u	J.Bassingthwaighte	Seattle	enzymes to body physiology
CellML,x+	Peter Hunter	U.Auckland	geometry, model sharing
GENESIS,u	James Bower	Caltech	neural networks
Simex,u+	Lael Gatewood	U.MN	Stochastic micro populations

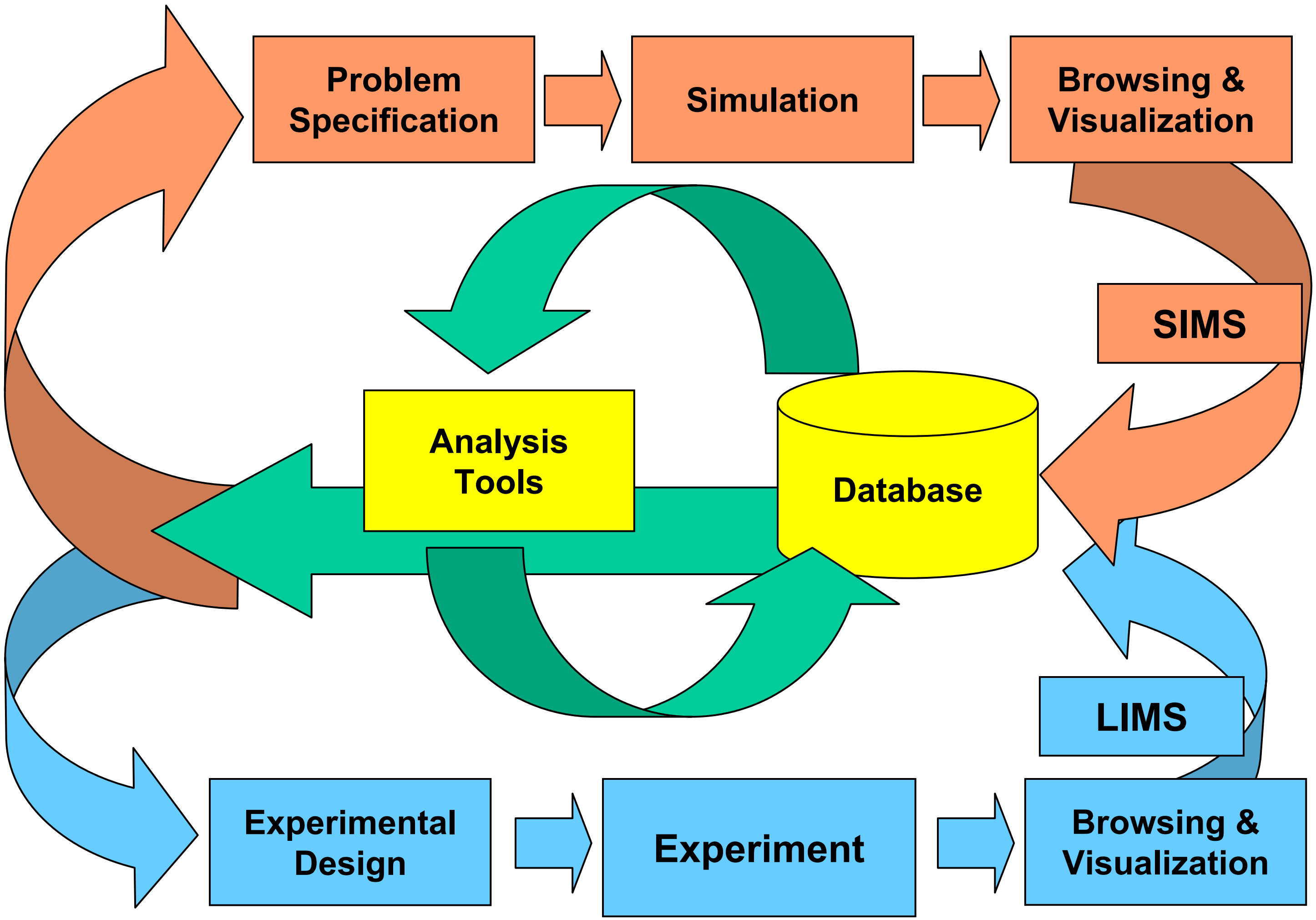
J=java, w = windows, u=unix, x=XML, + = source/community input

Visualization + Bioinformatics



Argonne Pathway Explorer on µMural Tiled Display

An Integrated View of Simulation, Experiment, and Bioinformatics

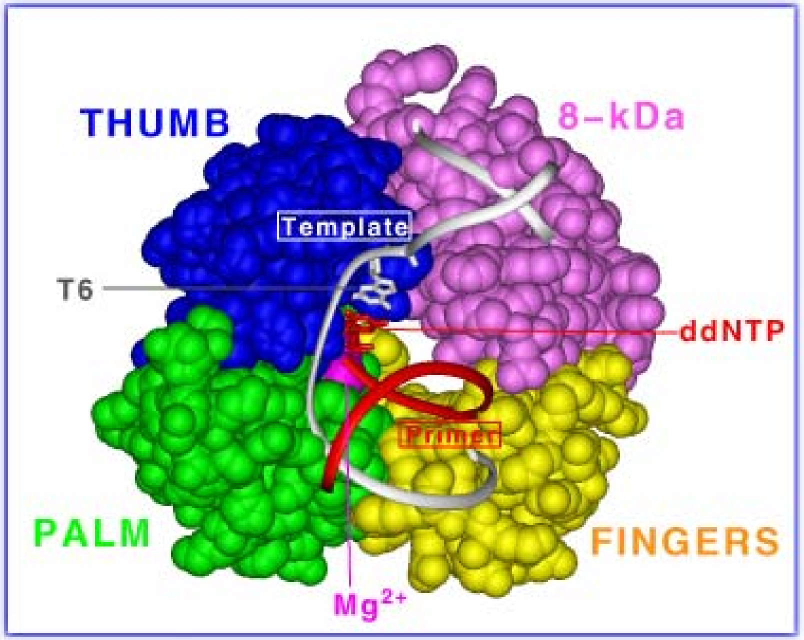


Samples of Current Projects In Biology from

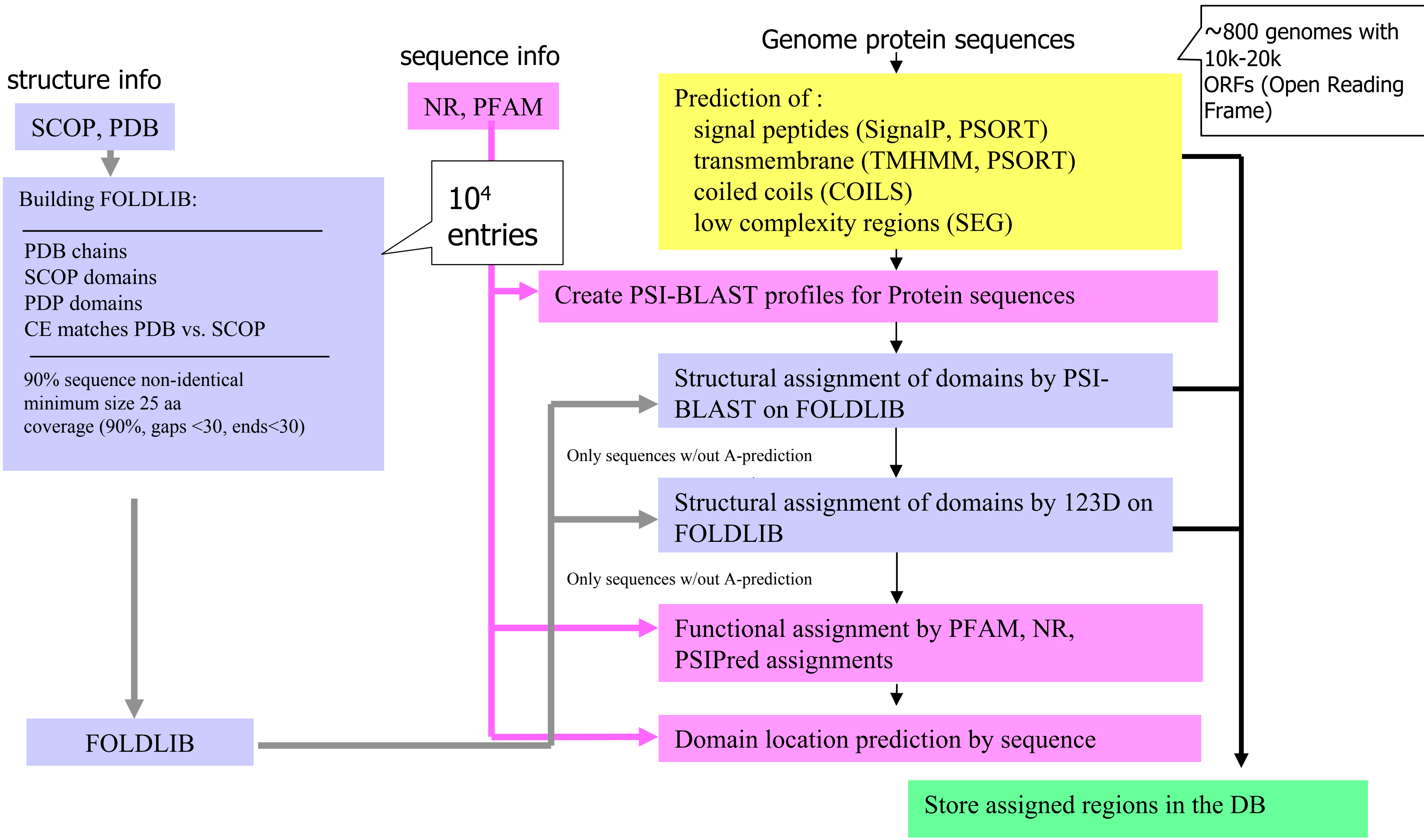
NCSA, SDSC, PSC, NERSC, and Utah

Tamar Schlick – NYU

- Cells have evolved sophisticated machinery to replicate and repair DNA accurately
- DNA polymerases – crucial components of this machinery – with hand-like subdomains
  - FINGERS - position DNA
  - PALM - phosphoryl transfer
  - THUMB - position incoming bp
- DNA synthesis error may play an important role in human aging and disease
- **Challenges:** probe at the atomic level, fidelity mechanisms employed to select correct nucleotide rather than the wrong one
- **Approach:** study large-scale conformational change that may help regulate synthesis fidelity



EOL Computational Pipeline

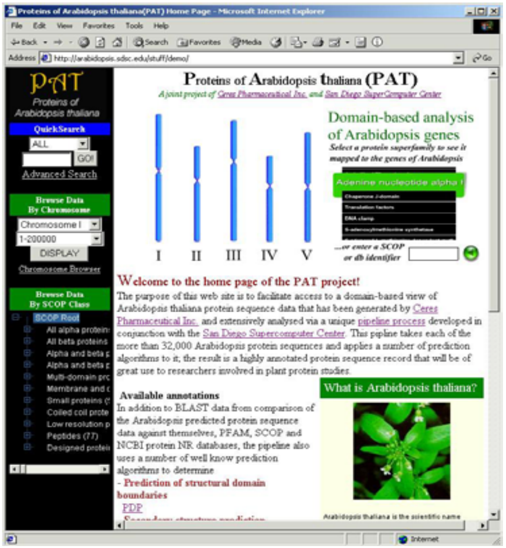


Encyclopedia of LIFE

"Almost everything in the body … is either made of proteins or made by them."

Matt Ridley, "Genome: The Autobiography of a Species in 23 Chapters"

- **Types of Questions which can be addressed by EOL**
  - ***Is protein X found in anthrax?***
  - ***Is protein X a drug target, that is, does it exist predominantly in pathogenic bacteria or is it found in eukaryotes also?***
  - ***Has caspase-1 (a protein involved in cell death and aging) been identified in any plants, if so what species and do the proposed protein structures look similar?***
  - ***Give me all available information on caspase-1***



Arabidopsis annotation

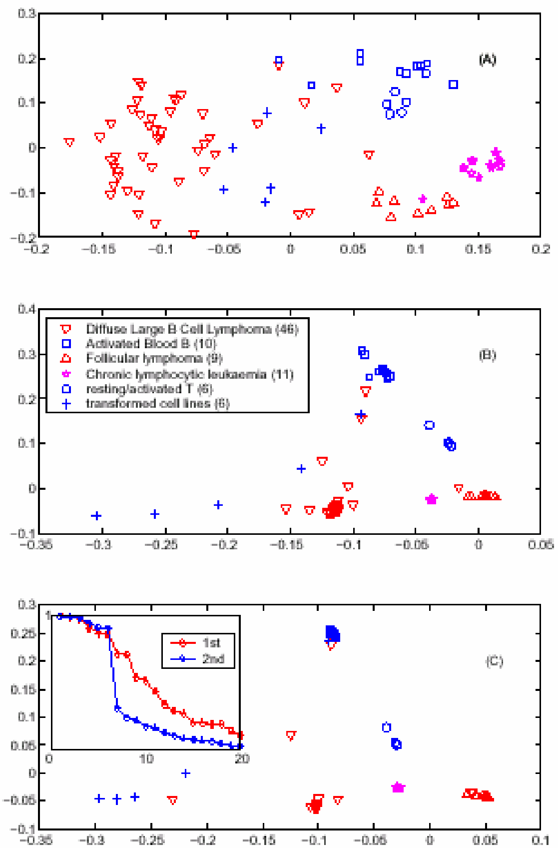
joint work with SDSC and Ceres

We've already started – *Annotation of* Arabidopsis thaliana *Proteins*

Data Mining for Genomics

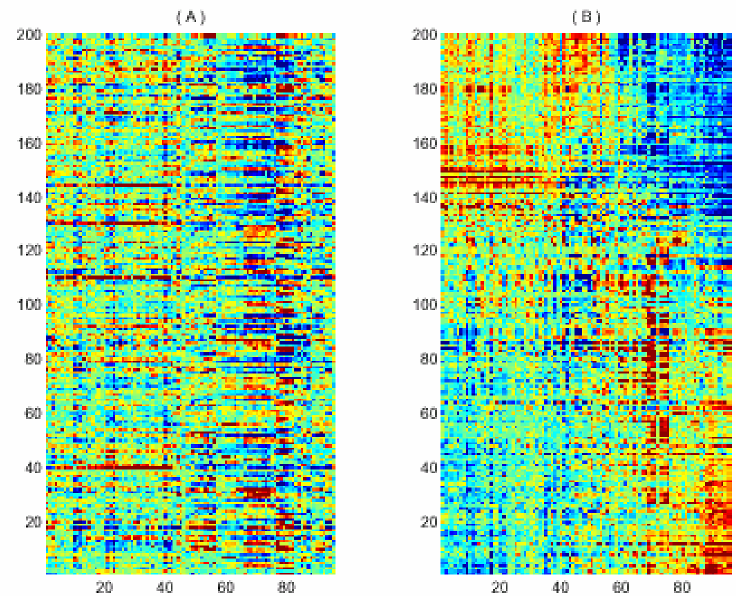
DNA Gene expression profiles

Class discovery/clustering

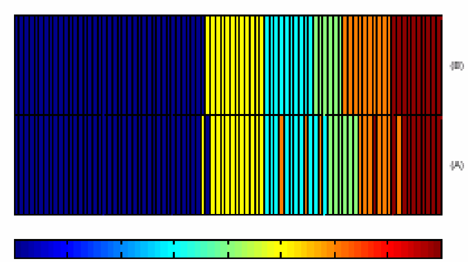


LDRD, Paper in RECOMB '02

Optimal ordering



Preserve cluster structure



Synaptic Transmission

Many neurological diseases due to problems of release or absorption of neurotransmitters like acetylcholine, glutamate, glycine, GABA, serotonin

Joel Stiles, PSC and Tom Bartol, Salk- MCell

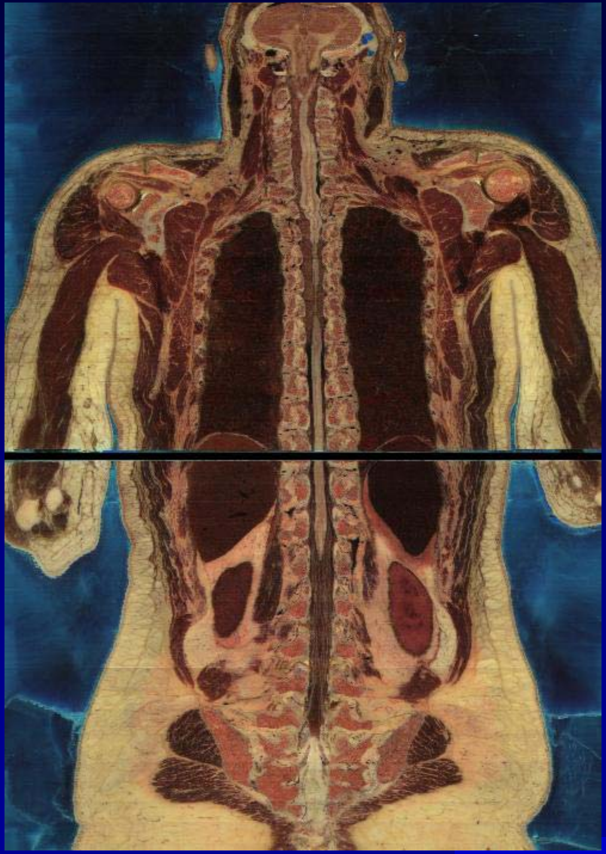
Unusual Medical Success

- In Slow Channel Congenital Myasthenic Syndrome, channel closes slower upon binding. Electrical current continues longer than normal.
- Particular patient presented puzzling symptoms
- Stiles experimented with the model parameters, and simulations showed that one could explain the symptoms if the receptors also opened slowly- then verified medically
- Unusual interplay of simulation and medical diagnosis- depends critically on realistic

geometry and on stochastic modeling

Teaching Anatomy-Networks and the Visible Human

- Permit multiple users in anatomy lab to fly through Visible Human in arbitrary directions (Network-sensitive)
- Use human perceptual factors, data compression to reduce bandwidth requirements

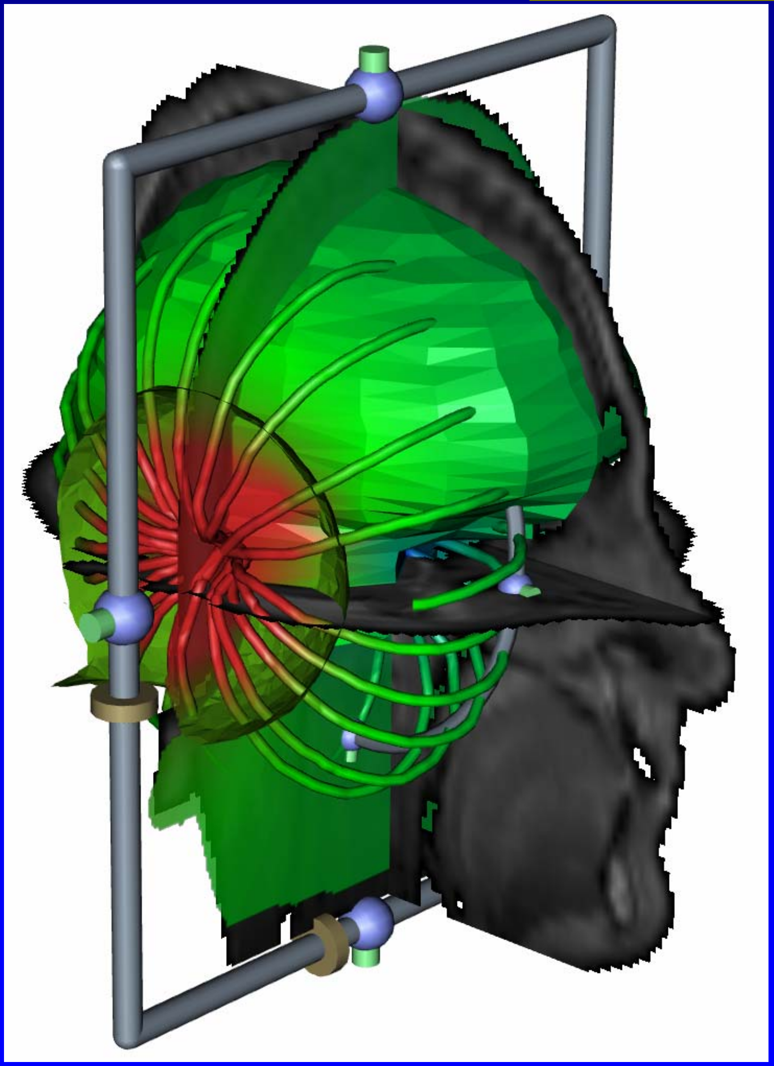


PITTSBURGH SUPERCOMPUTING CENTER

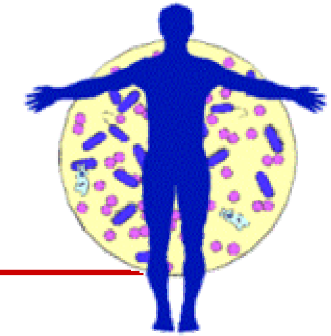
Time-critical: Neurosurgery

SCI Utah

Harvard & Brigham Women's Hospital



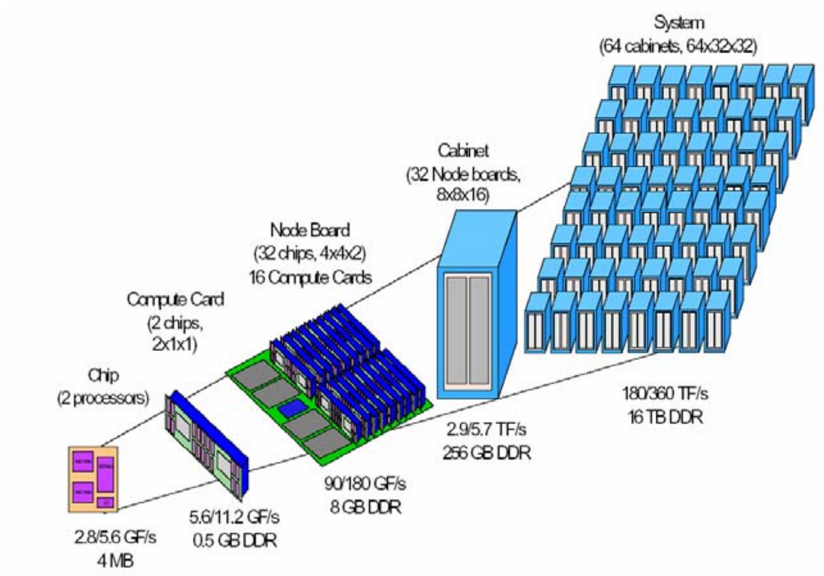
Future Vision

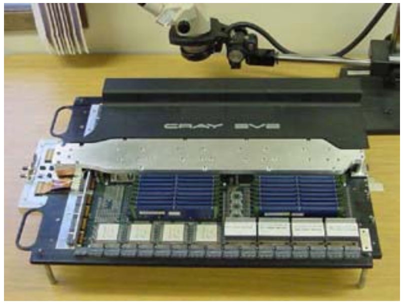


- Theory and Computation for Systems Biology
  - A focus on what makes things biological
- Integrated Modeling and Prototyping Tools
  - A Matlab for biological modeling
  - Portals and interfaces to existing simulation resources
- Integrated and Federated Databases
  - Frameworks and schema (e.g. discovery link, AfCS)
  - Xchange infrastructure (e.g. SBML, CellML, etc.)
- International "BioGrids" to Support Analysis, Modeling and Simulation
  - Beyond genomics and molecular modeling

Architecture Requirements for Biology

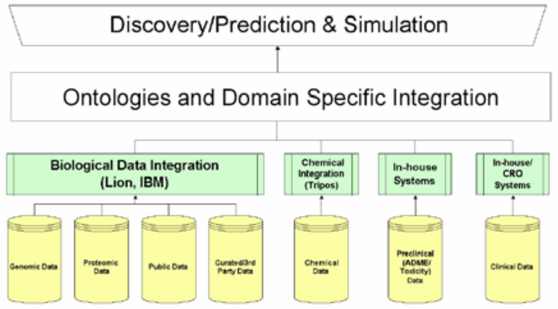
- Computational Biology is as Diverse as Biology itself
- Need for future systems
  - Capacity Computing
    - Clusters for high-throughput support
    - Automation of experimental laboratories
  - Capability Computing
    - Current: Protein science and Bioengineering
    - Future: cell modeling and virtual organisms
  - Data Intensive Computing
    - Data mining (genomes, expression data, imaging, etc.)
    - Annotation pipelines
  - Purpose built devices for well understood problems
    - Sequence analysis, imaging and perhaps protein folding





Grids and Biology

- Biology perhaps more than any other discipline can benefit from Grids
  - Biology is data intensive
  - Biology is highly distributed
  - Biology is moving very quickly
  - Biology is transitioning from an experimental science to a theory and computing driven science
  - Biologists are already dependent on the web
- The productivity gains from BioGrids will also directly impact drug development and delivery of medical care



BioGrid Services Model

Domain Oriented Services

- Drug Discovery
- Microbial Engineering
- Molecular Ecology
- Oncology Research

Basic BioGrid Services

- Integrated Databases
- Sequence Analysis
- Protein Interactions
- Cell Simulation

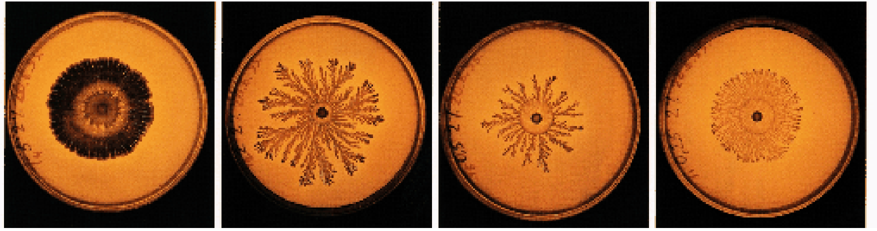
Grid Resource Services

- Compute Services
- Pipeline Services
- Data Archive Service
- Database Hosting

A Proposed International Systems Biology Grid

- A Data, Experiment and Simulation Grid Linking:
  - People [biologists, computer scientists, mathematicians, etc.]
  - Experimental systems [arrays, detectors, MS, MRI, EM, etc.]
  - Databases [data centers, curators, analysis servers]
  - Simulation Resources [supercomputers, visualization, desktops]
  - Discovery Resources [optimized search servers]
  - Education and Teaching Resources [classrooms, labs, etc.]
- Potentially finer grain than current Grid Projects
  - More laboratory integration [small laboratory interfaces]
  - Many participants will be experimentalists [workflow, visualization]
  - More diversity of data sources and databases [integration, federation]
  - More portals to simulation environments [ASP models]
- Global Grid Forum
  - Life Science Grid research group formed to investigate requirements
  - First meeting scheduled for GGF6 in Chicago mid October

Conclusions



- Biology is well positioned to co-dominate HPC applications for the next several decades
- Biological and Biomedical applications of HPC will require dramatic increases in both <u>capability computing</u> and <u>capacity computing</u>
- Data intensive computing is an important aspect of biological applications and will help drive high performance and high-function databases
- Biology and Grids are well suited for each other

Acknowledgements

- DOE, NSF, ANL, UC, Microsoft and IBM for support
- John Wooley (UCSD), Mike Colvin(LLNL/DOE), Richard Gardner (InCellico), Chris Johnson (Utah), Dan Reed (NCSA), Dick Crutcher (NCSA), Fran Berman (SDSC), Ralph Roskies (PSC), Horst Simon (NERSC) and others contributed to this talk

Backup Slides

Paths to Whole Cell Simulations

A = Algorithms
C = Compute
P = Parallelism
I = Integration

- Unregulated metabolic model (flux analysis)
- Allosteric regulation (binding changes conformation) (A)
- Gene Regulated + Metabolic Model (A, C)
- Heterogeneous/Compartmentalized/Diffusion (A,C,P)
- Active Regulation + Transport (A,C,P,I)
- Complete Integrated Cell (geometry) (A,C, P,I)

- Multicellular models (homogeneous) (P)
- Multicellular (homo) with complex communication (P)
- Multicellular (hetero) mixed population (P, I)
- Multicellular differentiation and motility (A, C, P, I)
- Multicellular structures with complex geometry $(A,C,P,I)^2$

Computer Science Barriers

- Framework for Functional Composability
  - Multiple modules
  - Multiple time scales and space scales
  - Empirical, semi-empirical, phenomological, data driven
- Interpretation of output of complex models
  - Visualization and automated interpretation
- Algorithms
  - Parameter estimation, graph theory, combinatorics
- Architectures and Software
  - Issues with scaling models and performance
  - Control and synchronization of multi component models